Marrying Relevance and Genre Rankings: an Exploratory Study*

Pavel Braslavski

Institute of Engineering Science RAS, Komsomolskaya 34,
620219 Ekaterinburg, Russia
pb@imach.uran.ru

Summary. In this chapter, we discuss different options for using genre-related information in Web search. We conduct an experiment on merging genre-related and text-relevance rankings using a reference Web collection. A method for automatic extraction of formality score akin to readability score using canonical discriminant analysis applied to a sample of genres with decreasing formality is proposed. Effects of aggregating genre-related and text relevance rankings are considered. Evaluation of the results shows moderate positive effects. Findings suggest that further research is needed on implicit use of genre-related information in Web search.

1 Introduction

Recent years have shown a growing interest to automatic genre analysis of Web documents, especially in the context of Web search. As the amount of indexed documents grows, the specification of a few keywords is not enough to describe user information need. Many studies suggest looking at document genre as an additional non-topical retrieval criterion. The output of a genre classifier could be used in Web search both explicitly and implicitly. Explicit use implies at least three possibilities. First, a focused ('vertical') search engine (SE) over documents belonging to a certain genre could be built. Second, the user can be given an opportunity to specify the desired genre in the query. Finally, the search engine results page (SERP) can be improved by enriching snippets with genre labels [2] or grouping the documents of the same genre together. However, all three options bring up issues.

* This paper expands the short paper presented at the workshop "Towards Genre-Enabled Search Engines: The Impact of NLP" [8].

[2] WEGA, a Firefox plug-in (see [31], Stein et al. in this volume), exemplifies this approach.

If we look at successful vertical search services such as scientific paper search[3], blog search[4], news search engines[5], or product search and comparison services[6] we notice that the task of gathering (or filtering out) content for services does not require especially sophisticated methods. Either the contributors are highly interested in providing their content to the service (scientific papers authors/publishers, on-line merchants), or the content is concentrated on several host sites in a certain form (like blog services, RSS feeds), or it can be found on the Web using simple surface features with high precision and satisfactory recall (e.g. scientific papers on authors' homepages).

Nowadays a simple search box and a ranked list of search results is a standard *de facto* for millions of search engine users. So the problem with the genre indicated explicitly in a query is that such advanced search option would be utilized by a marginal share of users.

The use of genre labels in search results presentation is somewhat questionable, too. Experiments [24] have shown that though most users expect genre information to be helpful for their Web search tasks, a straightforward implementation of genre-related hints does not improve user search effectiveness significantly. Moreover users can recognize distinct genres (such as catalog, FAQ, blog, or news) with high accuracy even from ordinary snippets [30].

The main problem in all three cases is to adapt or invent a suitable genre palette that is intuitively clear, complete, and unambiguous for the majority of users. Additionally, an appropriate interface needs to be designed. At present, this is just wishful thinking.

The approaches briefly described above imply an explicit use of genre within SEs. An alternative approach consists in using genre features in static (i.e. query-independent) ranking. Modern machine learning techniques allow for incorporating a wide variety of features to assess page quality regardless of the query (see [23] for an example). The features can be fairly heterogeneous and range from page popularity and pagerank to HTML well-formedness and color palette used. Some genres are less informative and convey mood or emotions rather than facts and information. One can try to incorporate this idea into the ranking scheme through machine learning. Another possible alternative is construction (as opposed to ranking) of SERP from documents of different genres. For example, if enough relevant documents are retrieved, there must be at least a news article, a product page, and a blog presented on the search results page. However, in this case too, one must decide an appropriate genre palette. A much more challenging task is to infer the genre from the query and return the documents of the implied genre to the user.

In this chapter, we suggest an additional alternative by incorporating genre information into relevance ranking.

[3] `http://citeseer.ist.psu.edu`, `http://scholar.google.com`
[4] `http://technorati.com`, `http://blog.yandex.ru`
[5] `http://news.google.com`, `http://news.yandex.ru`
[6] `http://shopping.yahoo.com`, `http://www.pricegrabber.com`

There are different definitions of *genre* or *style* (we treat these terms interchangeably). Both terms are widely used in linguistics, literary studies, aesthetics, art history and fashion. An extensive overview of different approaches to definition of *genre* can be found in [26]. We treat genre of a text document as a concept opposite to the document topic, similarly to several studies, e.g. [22]. We accept the intuitive understanding that genre is mainly related to the form (*how*) whereas the topic – to the content (*what*) of a document. This simplified approach is justified since we do not perform *genre classification/categorization*.

The idea of the experiment is to make use of a simple continuous measure of document's genre (akin to readability score). The approach is similar to static ranking in the way that we use a query-independent page-level feature in ranking, however we employ a much more straightforward approach – merging ranked documents. Although genre can be seen as 'orthogonal' to topic (i.e. almost all topics can be expressed in different genres), in the framework of our experiment we hypothesize that formal documents are potentially more informative than less formal ones.

We take a third-party system run from the *ad hoc* retrieval track at the Russian information retrieval evaluation seminar (ROMIP) provided partially with relevance judgments. Then we re-rank the documents according to genre-related score, merge both rankings with different weights, and compare the new ranks with the original ones using relevance judgments. The main unit of the analysis is an individual text-rich document. Due to specifics of the corpus used at the initial stage of the experiment we exploit only textual features in the analysis neglecting Web-specific genres and document features such as HTML markup and structure, and URL tokens.

We conducted our experiments on Russian documents but the methods can be easily applied to a different language.

In the part of genre-related score extraction the study is rooted in our early experiments on genre categorization [9]. In contrast to our previous study on genre 'admixture' in ranking [7] when we used an unsupervised approach, our current study employs a supervised method for extracting genre-related scores. The experiment is closely related to recent studies aimed at incorporating non-topical document facets into Web information retrieval (see Section 2).

The chapter is organized as follows. The next section surveys related work in the fields of genre classification, readability analysis, and previous experiments on incorporating genres into relevance ranking. Section 3 describes the data used in the experiment. Section 4 describes the extraction method and obtained formality score. Section 5 presents the produced rankings, evaluation metrics and final results. Section 6 concludes and outlines directions for future research.

2 Related Work

Our work is related to research in the fields of automatic genre classification, readability as well as information retrieval experiments on integrating genre-related features into relevance ranking schemata.

Genre classification.

After the pioneering work by Karlgren & Cutting [14] many papers on automatic genre classification have been published. The majority of the papers address the genre categorization problem and solve it using machine learning techniques. Set and structure of genre categories, corresponding learning sample, classification features, as well as learning technique constitute the diversity of the approaches. There are different sets of genres employed in the studies. The number of distinct genres ranges from binary classes (e.g. informative/imaginative; textual/non-textual) up to 16 multiple genre classes. Many researchers propose a hierarchical structure of genres. Some of them borrow an established set of genres, the others compile a genre palette based on a user study. The variety of classification techniques includes discriminant analysis, naïve Bayes, logistic regression, neural networks, kNN, and SVM. A number of studies utilize existing corpora for learning and evaluation, the rest compile their own. As opposed to topical text categorization most genre categorization studies use mainly non-content features such as surface text statistics, function words count, POS and punctuation mark frequencies, etc. For more details one can refer to a comprehensive survey of the field [27]. At least two noteworthy papers appeared after the survey that primarily concentrate on analysis of Web documents. Meyer zu Eissen & Stein [19] conducted a user study spawning a set of eight Web genres useful for Web search, and built a corpus containing these genres (the KI-04 corpus). Along with linguistic features traditionally used in genre analysis, their study employs HTML-based features. Lim et al. [17] expanded this approach even further and made use of a wider range of features (326 in total), including various surface, lexical, syntactic, HTML, and URL features.

Automatic genre analysis is not restricted to genre categorization – there are some efforts on genre clustering. Rauber & Müller-Kögler [22] adapted an unsupervised technique for revealing genre-dependent similarities between documents. The self-organizing map (SOM) was used to cluster documents according to their various surface level text features. The results of analysis were incorporated into a content-based representation through coloring individual documents according to their location on the resulting SOM. Gupta et al. [12] applied the notion of Web site genre to improve web page cleansing methods (i.e. removal of ads, unnecessary images and extraneous links). Sites are clustered in word feature space using cityblock distance. The distinction of the method is that sites are characterized not only by the words they contain

but also by the words from snippets returned by several SEs in response to the web site domain name.

Readability scores.

Research on readability has its roots in psycholinguistics but in fact is very similar to automatic genre analysis. The aim is to obtain a simple measure to compare the comprehension complexity of texts conveying similar meaning using surface cues [11].

The 'traditional' way to construct a readability formula is as follows. First, text complexity estimates are obtained experimentally. Second, text features that potentially contribute to its complexity are extracted. Third, text features and text complexity are tied together using regression analysis. There are different psycholinguistic techniques to measure text complexity: reading time (normalized by the individual reading skills), post-reading questionnaires assessing text comprehension, and cloze tests. There are different features used in readability formulæ: number of words from different word lists (such as 'easy', 'hard', 'abstract', 'most frequent', etc. word lists), word length, sentence length, number of sentences per paragraph, number of prepositional phrases, etc. In summary, all the features can be divided in two classes: semantic features reflecting the complexity of vocabulary and structural features reflecting compositional complexity (usually on the sentence level, sometimes on the paragraph level).

There have been recent papers introducing a novel approach to readability that is very close to ours.

Si & Callan [28] and Collins-Thompson & Callan [10] re-formulate the readability prediction task as a categorization problem: they use labeled data (documents with assigned readability labels), tokens as features, and naïve Bayes classifier. Their approach emphasizes semantic features, i.e. difficulty of a text is defined entirely through its vocabulary. The method outperforms traditional readability measures on Web data.

A related study is described in [16]: a query-independent 'familiarity classifier' is build upon several hundreds of documents manually tagged as 'introductory' or 'advanced' using random forest classifier. Three groups of feature are employed: 1) stop-words, 2) common readability features and traditional readability scores themselves, 3) features based on various characteristics of web page documents (e.g. anchor text count or Similarity of WordNet expansion of top 10% of document with remaining 90%). The authors show that traditional readability measures such as Fog index, Flesch readability score, and Flesch-Kincaid grade level capture the notion of familiarity poorly. However, the method does not consider topic relevance: top-20 documents returned by a search engine are all assumed to be relevant to the query, which seems to be a very strong assumption.

Genres in relevance ranking.

Strzalkowski et al. performed stylistic analysis on TREC data already in 1995 [29]. Their idea was to find stylistic features that could discriminate relevant and non-relevant documents. Using previous TREC results, they found that relevant documents tend to be more complex on different levels - textual, syntactic, and lexical. A decision tree classifier was built upon labeled data, documents classified as non-relevant were to move to the end of the list. However this strategy did not gain in average precision: "The consequence is that to make use of stylistic variation for reliable relevance grading we need a query typology: each query must be identified for likely style preferences."[29] As a matter of fact, our study reasserts these findings.

A High Accuracy Retrieval from Documents (HARD) track was organized within TREC campaign in 2003-2005 [3, 4]. The goal of the track was "to bring the user out of hiding, making him or her an integral part of both the search process and the evaluation" [3] as opposed to an abstract 'average' user behind traditional TREC topics. TREC topics were provided with metadata including GENRE and FAMILIARITY items. In particular, in HARD 2004 track GENRE had values of *news-report*, *opinion-editorial*, *other*, or *any*; FAMILIARITY had a value of *little* or *much*. Within HARD track RELEVANT judgment means that the document is on topic *and* it satisfies the appropriate metadata. Attempts to utilize the available metadata, including GENRE and FAMILIARITY are exemplified by track reports [1, 6]. Belkin et al. [6] used readability scores, average number of syllables per word, and abstractness/concreteness of the document's vocabulary to model familiarity. Genres were modeled by language models; the Kullback-Leibler (KL) divergence determined whether a document was a member of the genre. Final rankings were obtained via weighted combination of baseline scores and metadata classifiers' scores. Both genre and familiarity classifiers performed poorly. As the authors stated, "using language models to capture genre preference was a complete failure, presumably because the language models captured the topics of the training documents."

Abdul-Jaleel et al. [1] were more successful at building genre classifier. They used linear SVM and 10K most frequent tokens in the corpus, subcollection tags, and various length measures of a document as features. Final rankings were produced by linear combination of the normalized outputs of both the retrieval and classifier outputs. Although genre classifier showed good performance, it did not leverage the retrieval effectiveness. Authors noticed that "many documents judged relevant clearly fall outside the requested metadata. Searchers know a relevant document when they see one, but a priori they do not fully know what metadata is required of a relevant document."

In the following sections, we will describe experiments that complement the approaches described above.

3 Data

In this study we use two datasets of Russian documents: 1) a small corpus of five functional styles as a learning sample for extracting a genre-related score and 2) a subset of reference ROMIP Web collection for experimentation and evaluation purposes.

3.1 Functional Styles Sample

For our experiment we needed a simple measure that captured the formality or 'seriousness' of the document akin to a text readability measure. Unfortunately, there is no widely accepted and use-proven readability index for Russian. For the purpose of obtaining a genre-related score we reused a functional styles sample that was employed in our previous experiments. The sample contains 50 federal acts (official functional style), 54 scientific papers in natural sciences (academic style), 61 online news articles (journalistic style), 79 short stories by modern Russian authors (literary style), and 61 fragments of online chats (everyday communication style) – 305 documents in Russian in total.

It is important to stress that our study is not aimed at building a functional styles classifier. The assumption is that formality progressively decreases from federal acts to chats, being federal act the most formal genre and chat the least formal.

3.2 ROMIP Collection

ROMIP stands for Russian Information Retrieval Evaluation Seminar which is a Russian TREC-like information retrieval evaluation initiative [25]. ROMIP Web collection contains about 600,000 HTML pages in Russian from the free Web hosting `narod.ru` and adequately reflects the diversity of Web genres. The collection is used in the ROMIP ad hoc retrieval track and is freely available upon request.

Along the documents the collection contains a list of about 20,000 queries taken from a real-life Web SE query log. Each participating system performs the whole set of queries over ROMIP collection. A small selection of queries (or topics in TREC terminology) is evaluated manually using a pooling method in each yearly cycle. A short description (close to TREC's *narrative*) representing one of the possible query interpretations is provided to help assessors (Fig. 1). Many descriptions imply detailed and informative documents. This fact suggests that ranking 'serious' documents higher may improve the overall search quality within the ROMIP framework. We implement this approach in our experimental framework, however it will not comply with all real-life information needs obviously.

In our previous stylistic experiments [7] we found out that menus, navigation, ads, authorship and copyright notices, etc. presented on the majority

Query arw13494: memory training
*Description*Documents containing advice for human memory improvement, diverse techniques for memory training. Documents containing recipes of food supplements are useful. Especially important are documents containing detailed and precise instructions for those who want to train their memory.

Query arw19003: are we alone in the universe?
Description: The page must contain information on extraterrestrial intelligence research, existing hypotheses as well as different opinions on this issue.

Query arw18885: why do the airplanes fly
Description The page must contain information about airplanes, aerodynamics basics, wing lift.

Fig. 1. Sample ROMIP topics: query and its description (originally in Russian, descriptions are used on the evaluation stage only).

of HTML pages in the ROMIP collection significantly skew genre-related parameters. So we took the collection after template removal routine described in [2]. However, the difference from the original collection was not substantial since the ROMIP collection is compiled from free hosting pages and includes mainly sites with moderate number of pages which makes proper template detection and removal difficult. All documents were converted to Windows-1251 Cyrillic encoding and subsequently to *plain text*.

For our experiment, we took the results of one of the ROMIP'2006 participating systems which utilizes only text relevance features [13]:

- single query terms match;
- pairs of query terms match;
- exact phrasal match;
- all query terms appear in the document;
- a significant part of query terms appears within a sentence.

Additionally, pseudo-relevance feedback techniques were used. The system was trained on relevance judgments of two previous campaigns – ROMIP'2004 and ROMIP'2005.

This ROMIP subset contains 6,906 documents corresponding to 70 evaluated search topics (67 topics with 100 ranked documents per topic plus three topics with 23, 87, and 96 documents, respectively). The majority of these documents have binary relevance judgments: 5393 documents (420 relevant + 4973 non-relevant) with so-called 'strong' judgments (i.e. all assessors agreed on judgment) and 5416 documents (1105 relevant + 4311 non-relevant) with 'weak' relevance judgments (i.e. at least one assessor judged a document as 'relevant'). Some topics have no corresponding relevant documents (13 in case of 'strong' relevance and three in case of 'weak' relevance). The rest of the documents have tag 'can't be judged' or do not fall into the evaluated document pool. The pool depth in ROMIP'2006 was 50, i.e. the first 50 documents from the participating systems' runs were pooled and evaluated. At the 50 cut-off

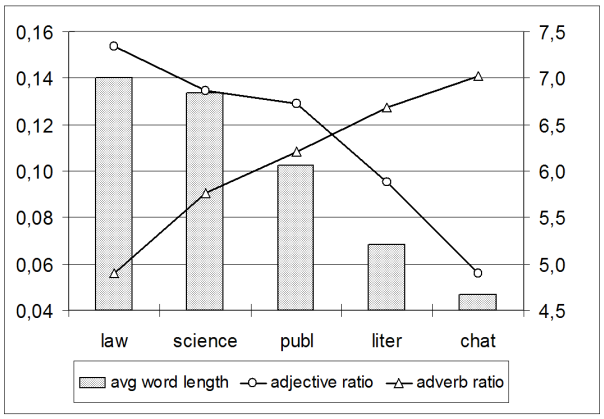


Fig. 2. Selected characteristics of the functional styles sample.

the statistics of the subset looks as follows: 3473 documents in total, including 354 and 899 relevant documents (strong and weak judgments, respectively); topics with zero relevant documents – 15 and 4 (strong and weak judgments, respectively).

4 Formality Score

As we mentioned before, there is no widely accepted and use-proven readability score for Russian that would be appropriate for our aims. So we opted for building a 'formality score' based on our previous research.

In our earlier experiments on genre categorization [9] we employed the concept of functional styles, which is well-established in Russian linguistics. There are five basic functional styles: *official*, *academic*, *journalistic*, *literary* (*fiction*, *belles-lettres*), and *everyday communication style*. Functional styles have been the subject of an study on automatic stylistic analysis [20]. More details on the theory of functional styles can be found in [15].

Our approach is rather operational. We consider five functional styles simply as text classes of gradually decreasing formality. We use this small sample only for building a genre-related score and then 'throw this ladder away after climbed up it'. The quantitative characteristics of the functional styles sample confirm appropriateness of the approach (Fig. 2). Such features as average word length (one of the most commonly used features in different readability formulae) and POS distribution change monotonically over five styles.

We use *canonical discriminant analysis* to extract the formality score. The method is illustrated in Fig. 3: feature space transformation is performed in order to find a direction (a weighted sum of initial features) with the best separating ability between classes. The method is similar to *principal components analysis* (PCA); the difference is that class structure is taken into

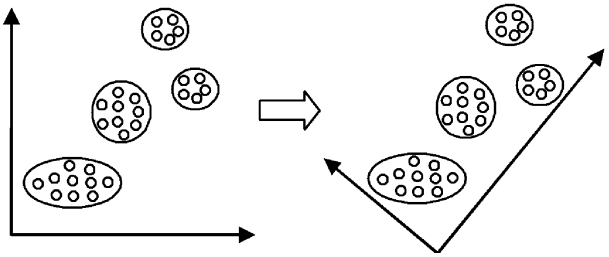


Fig. 3. The idea of canonical discriminant analysis.

account. After this point we abandon the set of discrete genres and proceed to an continuous index.

We experimented with different genre-related easily computable textual features. As mentioned before we use exclusively real-value textual features implying relatively long (since all features are averages) and coherent text documents. The features we used are quite similar to those used in our previous experiments and other genre analysis studies: surface features like word and sentence length (the latter is based on a simple rule for sentence boundary detection), punctuation and functional words counts, and POS ratios (using *mystem* POS tagger [21] without any disambiguation). Feature selection process was guided by the percentage of explained variance, analysis of variance over five classes, as well as considerations on feature semantics. After a series of trials we opted for a combination of nine features. The formula for the first canonical root that we treat as formality score is as follows (standardized values, greater values correspond to lesser formality):

$$S = -0.49x_1 + 0.27x_2 + 0.46x_3 + 0.04x_4 + 0.24x_5 + 0.32x_6 - 0.48x_7 + 0.32x_8 - 0.11x_9,$$

where

x_1 – average word length;
x_2 – smiley count;
x_3 – finite verb count;
x_4 – adjective count;
x_5 – first person pronoun count;
x_6 – expressive punctuation count;
x_7 – neuter noun count;
x_8 – adverb count;
x_9 – genitive chain count.

The first canonical root explains 84% of sample's variance. Fig. 4 shows that although the classes are not smoothly separable in this 2D space, they line up along X axis, preserving their 'formality order' in general.

The obtained index is fairly similar to a readability score: average word length, a component of almost all readability measures, enters into the formula with negative weight, the same way as genitive chain count (reflects syntactic complexity) and neuter noun count (neuter nouns tend to be more abstract in

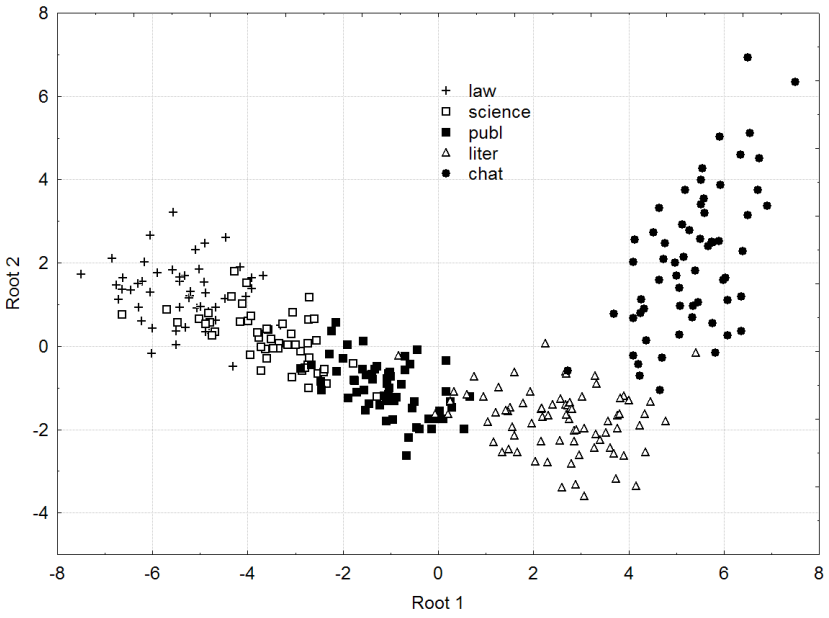


Fig. 4. Scatter-plot of the learning sample in the 1^{st} and 2^{nd} canonical roots.

Russian). In contrast smiley, expressive punctuation and first person counts enter into the formula with positive weights, reflecting text informal flavor. For convenience we mapped the obtained canonical root onto [0, 1] interval with lesser values corresponding to lesser formality.

The applied corpus-based approach is low-cost, flexible, and easily adjustable compared to traditional methods for building readability scores based on reading tests. A thorough examination of the obtained index and comparison with the existing readability indices will be addressed in a separate study.

5 Results

5.1 Genre-Related Rankings

We calculated formality scores for documents in our ROMIP subset. We performed a selective comparison of documents and can estimate that obtained index reflects formality perception accurately. Relevant documents appeared to be somewhat 'more formal': averaged formality scores for our ROMIP subset are 0.62 and 0.59 for relevant and non-relevant documents, respectively (the difference is significant at $p < 0.005$). Distribution of formality score values over Web documents sample is presented on Fig. 5. One can see that distribution is fairly smooth, 'neutral' documents constitute the majority of the sample.

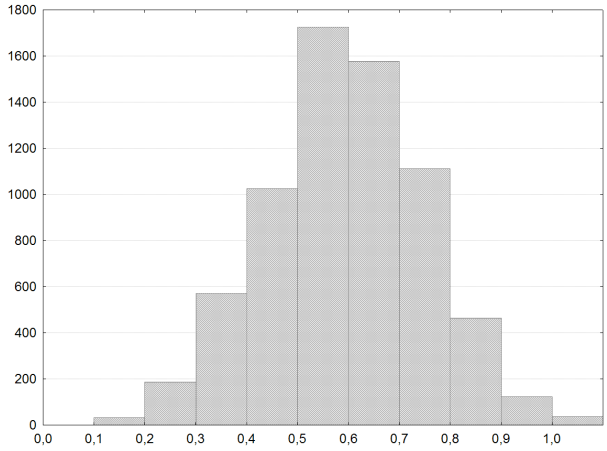


Fig. 5. Distribution of mapped formality values over ROMIP sample (6848 documents).

The obtained formality score similarly to readability indices implies coherent text. There are many types of non-textual web documents such as link and price lists, input forms, photo galleries, home pages with predominantly presentational content, etc. In order to filter out such documents as far as possible using simple methods, we introduced two restrictions for documents to be re-ranked: 1) longer than five sentences and 2) finite verb/sentence ratio greater than threshold (a simple signal of text coherence, threshold is selected empirically).

All evaluated documents meeting the restrictions have been ranked according to the genre-related score in descending order within each topic (more 'formal' documents on the top); all other documents preserved their initial positions. We obtained four initial genre-related rankings:

1. **T100**: all documents longer than five sentences (4846 documents processed);
2. **T100C**: additionally, finite verb/sentence ratio ≥ 0.6 (3823 documents processed);
3. **T50**: top-50 documents in each topic longer than five sentences (3030 documents processed);
4. **T50C**: additionally, finite verb/sentence ratio ≥ 0.6 (2332 documents processed).

In the next step, we aggregated the obtained genre-related ranks (R_G) with the initial keyword-relevance ranks (R_Y) (see [13] for details on R_Y). We used a straightforward approach to aggregation: new rank was computed based on a linear combination of text relevance and genre-related ranks, i.e. $R_Y + \alpha R_G$ (α is weight of genre-related ranking, $\alpha \in [0, 1]$). This scheme can

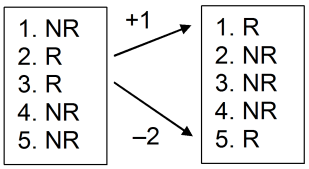


Fig. 6. Rank displacement of relevant (R) documents (for this example $D_R = -1$).

be referred to as a simple case of weighted Borda method that is widely used in different areas, including rank aggregation for metasearch. It is important to note that we did not aim at finding an optimal α for the rank combination. Although the number of processed documents is appreciable, the number of topics with relevant documents does not allow us to test our results properly and generalize well. The proposed re-ranking method is fairly conservative. Apart from the fact that many short and presumably incoherent documents preserve their positions since we are not confident enough to assign them a formality score, small α values prevent documents from distant jumps.

For evaluation of the aggregated ranks we use *rank displacement of relevant documents* (D_R) – a metric introduced in [5] for evaluation of data fusion effects in information retrieval. D_R sums the ups and downs of relevant documents in the new list in comparison to the original one (Fig. 5). Note that small movements in the top of the list 'cost' the same as in the bottom. Furthermore, we count up absolute number of tasks with positive and negative values of D_R. Additionally, we use official ROMIP metrics: *mean average precision* (*MAP*, calculated for the top-50 documents), *p1*, and *p10* (precision at levels 1 and 10, respectively). Note, that average precision (AP) is highly sensitive to ranking of relevant documents in contrast to D_R, thus little movements of relevant documents in the bottom of the ranked list have almost no effect on this metric; conversely small drops of relevant documents in the top of the list impair the metric value significantly.

5.2 Merged Rankings

The most illustrative results are obtained on weak relevance judgments. Fig. 7–10 show both macro- and micro-averaged D_R values and absolute numbers of topics with positive vs. negative changes depending on genre-related rank's weight for T100, T100C, T50, and T50C rankings. Standard ROMIP metrics for T100C and T50C rankings are shown in Fig. 11.

As one can see a small admixture of genre-related scores can slightly improve relevance ranking in terms of D_R metric. As Fig. 8 shows, in the best case approximately every second relevant document in each topic climbs one position higher in average. A simple criteria for text coherence based on finite verb ratio increases maximum macro-averaged D_R and broadens the range of its positive values and at the same time flattens the difference between topics with positive and negative effects. In case of the pool-deep re-ranking (T50C)

the use of this criteria keeps the macro-averaged D_R values in the positive half-plane and positive changes majorize negative changes at topic level.

However, these positive effects are not reflected in the standard ROMIP metrics except for an insignificant growth of *MAP* (less than 1%, Fig. 11(b)) and some occasional splashes on *p1* plot (Fig. 11). Fig. 12 illustrates difference in average precision between initial ranking ($\alpha = 0$) and merged one ($\alpha = 0.2$) by topic (the outlying topic is *arw13494: memory training*).

If we take a look at individual topics, we can find approximately 25 of them that are responsive to mixing genre ranks with traditional keyword-relevance ranks in almost every proportion. The examples are (originally in Russian):

- *arw17563: what to feed a cat on*
- *arw2000: meals in the fast*
- *arw5608: quantum computer*
- *arw10947: tv commercial creation*
- *arw2755: all about al capone*

We were unable to find a reliable pattern for these topics based on mean and standard deviation of the formality score, number of relevant documents, etc. According to the subjective observation descriptions of these topics might represent a more rigorous interpretation within ROMIP evaluation than a common one. But at the same time, mean formality score designates 'serious' topics with confidence. For example, these five topics with maximum mean formality scores consist mainly of legal, financial, medical, and popular scientific documents (originally in Russian):

- *arw12162: contract-based [military] service*
- *arw2538: magnetic field effects on humans*
- *arw18557: harmful effects of polluted air on respiratory apparatus*
- *arw16263: what is a promissory note*
- *arw7927: national income*

6 Conclusion

In this chapter we investigated different options for using genre-related information in Web search and consider the implicit use of such information most promising.

We conducted an experiment on merging genre-related and text-relevance rankings using reference ROMIP Web collection. To this end we proposed a method for automatic extraction of formality score using canonical discriminant analysis applied to a small sample of functional styles. Evaluation of the aggregated ranks shows that we can achieve moderate improvements on our experimental data set in average by mixing in a small fraction of genre-related rank. Notably, there is a subset of queries that is quite responsive to mixing

genre ranks with traditional keyword-relevance ranks. These findings confirm previous results on incorporating genre information into relevance ranking.

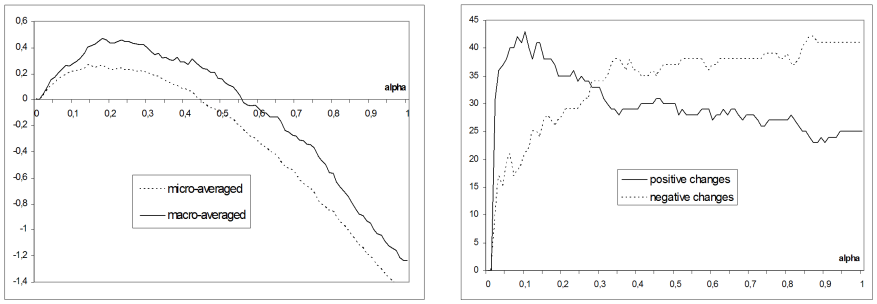
Our study suggests that a promising direction for future research could be incorporating genre information into static ranking. To this end a comprehensive study of distinctive genres' usefulness has to be carried out.

Another possible direction could be inferring of the expected genre (or *genre range* when thinking of continuous genre index) of the answer based on query processing. To the best of our knowledge the sole study on predicting user's education level based on a query is paper by Liu et al. [18]. The study demonstrates good quality in classifying queries according to student grade. The approach uses SVM and various features derived solely from queries, including sentence and word length features, percentage of part-of-speech tags, various readability indices, as well as frequency of numerous 1-, 2-, and 3-word sequences. Yet the paper deals with natural language questions rather than real Web SE queries and the problem remains open. A more reliable way could be click data analysis for frequent queries in order to estimate most expected document genres for those queries.

A further option could be accounting for genres in the personalized search framework. The problem is that a user's genre expectations vary from topic to topic, and drift unevenly with time.

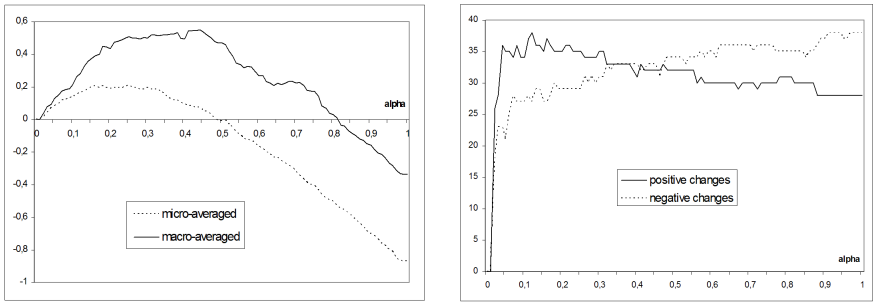
Acknowledgements.

We would like to thank Mikhail Ageev and Andrei Tselishchev for their help with data processing. We also thank Yandex (`www.yandex.ru`) for providing us with the experimental data. Many thanks to Matthew McCool and volume editors for their valuable comments on the draft.



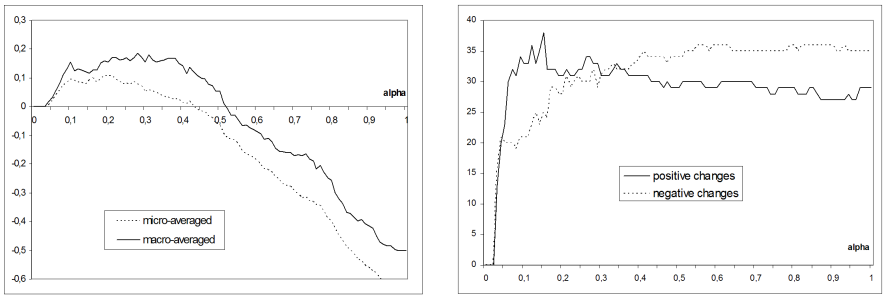
(a) (b)

Fig. 7. T100 re-ranking results: (a) averaged rank displacement; (b) number of tasks with positive and negative D_R.



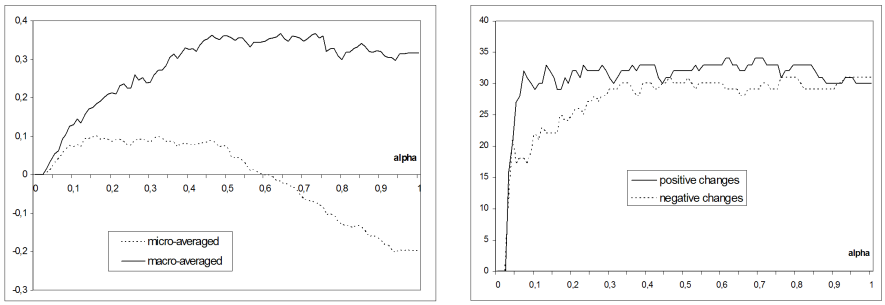
(a) (b)

Fig. 8. T100C re-ranking results: (a) averaged rank displacement; (b) number of tasks with positive and negative D_R.



(a) (b)

Fig. 9. T50 re-ranking results: (a) averaged rank displacement; (b) number of tasks with positive and negative D_R.



(a) (b)

Fig. 10. T50C re-ranking results: (a) averaged rank displacement; (b) number of tasks with positive and negative D_R.

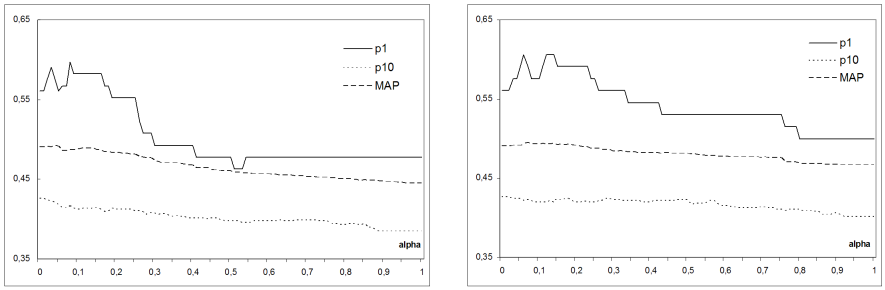


Fig. 11. Standard ROMIP metrics: (a) T100C; (b) T50C.

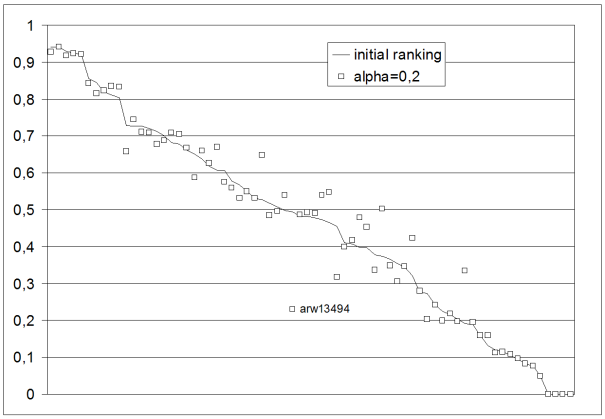


Fig. 12. Initial vs. new ($\alpha = 0.2$) average precision (T50C), topics sorted by initial AP.

References

[1] Abdul-Jaleel, N., Allan, J., Croft, W.B., Diaz, F., Larkey, L., Li, X., Smucker, M.D., Wade, C.: UMass at TREC 2004: Novelty and HARD. In Proceedings of TREC 2004 (2005)

[2] Ageev, M., Vershinnikov, I., Dobrov B.: Extraction of the significant Part of Web Pages for Information Retrieval (in Russian) [Izvlečenie značimoi informacii iz web-stranic dlja zada informacionnogo poiska]. In: Internet-Matematika, pp. 283–301 (2005) Available online: `http://company.yandex.ru/grant/2005/07\_Ageev\_102942.pdf`

[3] Allan J.: HARD Track Overview in TREC 2003: High Accuracy Retrieval from Documents. In: Proceedings of TREC-2003, pp. 24–37 (2004)

[4] Allan J.: HARD Track Overview in TREC 2004: High Accuracy Retrieval from Documents. In: Proceedings of TREC-2004, pp. 25–35 (2005)

[5] Beitzel, S.M., Jensen, E.C., Chowdhury, A., Grossman, D., Frieder, O., Goharian, N.: Fusion of Effective Retrieval Strategies in the Same Information Retrieval System. JASIST, 55(10), 859–868 (2004)
[6] Belkin, N., Chaleva, I., Cole, M., Li, Y.-L., Liu, L., Liu, Y.-H., Muresan, G., Smith, C., Sun, Y., Yuan, X.-J., Zhang, X.-M.: Rutgers' HARD Track Experiences at TREC 2004. In: Proceedings of TREC-2004 (2005)
[7] Braslavski, P., Tselishchev, A.: Style-Dependent Document Ranking. In: Proceedings of the 7th Russian Conference on Digital Libraries (RCDL'2005), pp. 159–164 (2005) Available online: `http://www.rcdl2005.uniyar.ac.ru/ru/RCDL2005/papers/sek7\_1\_paper.pdf`
[8] Braslavski, P.: Combining Relevance and Genre-Related Rankings: an Exploratory Study. In: Proceedings of the International Workshop "Towards Genre-Enabled Search Engines: The Impact of NLP", Borovets, Bulgaria, pp. 1–4 (2007) Available online: `http://kansas.ru/pb/paper/ranlp2007.pdf`
[9] Braslavski, P.: Document Style Recognition Using Shallow Statistical Analysis. In: Proceedings of the ESSLLI 2004 Workshop on Combining Shallow and Deep Processing for NLP, Nancy, France, pp. 1–9 (2004) Available online: `http://esslli2004.loria.fr/content/readers/36.pdf`
[10] Collins-Thompson, K., Callan, J.P.: A Language Modeling Approach to Predicting Reading Difficulty. In: Proceedings of HLT/NAACL, pp. 193–200 (2004)
[11] DuBay, W.H.: The Principles of Readability (2004) Available nline: `http://www.nald.ca/fulltext/readab/readab.pdf`
[12] Gupta, S., Kaiser, G., Stolfo, S., Becker, H.: Genre Classification of Websites Using Search Engine Snippets. In: Proceedings of SIGIR'2005 Workshop "Stylistic Analysis Of Text For Information Access", Salvador, Bahia, Brazil (2005)
[13] Gulin, A., Maslov, M., Segalovich, I.: Yandex' Algorithm for Text Relevance Ranking at ROMIP'2006 (in Russian) [Algoritm tekstovogo ranžirovanija Jandeksa na ROMIP'2006]. In: Proceedings of ROMIP'2006, Suzdal, Russia, pp. 40–51 (2006) Available online: `http://www.romip.ru/romip2006/03\_yandex.pdf`
[14] Karlgren, J., Cutting, D.: Recognizing Text Genres with Simple Metrics Using Discriminant Analysis. In: Proceedings of the 15th Conference on Computational Linguistics, pp. 1071–1075 (1994)
[15] Kožina, M.N.: Foundations of the Functional Stylistics (in Russian) [K osnovaniyam funkcional'noi stilistiki], Perm (1968)
[16] Kumaran, G., Jones, R., Madani, O.: Biasing Web Search Results for Topic Familiarity. In: Proceedings of CIKM'05, pp. 271–272 (2005)
[17] Lim, Ch.S., Lee K.J., Kim G.Ch.: Multiple Sets of Features for Automatic Genre Classification of Web Documents. Information Processing and Management 41, pp. 1263–1276 (2005)

[18] Liu, X., Croft, W. B., Oh, P., Hart, D.: Automatic Recognition of Reading Levels from User Queries. In: Proceedings of SIGIR'2004, pp. 548–549 (2004)

[19] Meyer zu Eissen, S., Stein, B.: Genre Classification of Web Pages. In: Proceedings of the 27th German Conference on Artificial Intelligence (KI-2004), Ulm, Germany, pp. 256–269 (2004)

[20] Michos, S., Stamatatos, E., Fakotakis, N., Kokkinakis, G.: Categorizing Texts by Using a Three Level Functional Style Description. In: Rasmsay, A.M. (ed.) Artificial Intelligence: Methodology, Systems, Applications, Frontiers in Artificial Intelligence and Applications, Vol. 35 (1996) Available online: http://slt.wcl.ee.upatras.gr/papers/michos2.pdf

[21] Mystem tool, http://company.yandex.ru/technology/products/mystem/mystem.xml

[22] Rauber, A., Müller-Kögler, A.: Integrating Automatic Genre Analysis into Digital Libraries. In: Proceedings of the JCDL'2001, pp. 1–10 (2001)

[23] Richardson, M., Prakash, A., Brill, E.: Beyond PageRank: Machine Learning for Static Ranking. In: Proceedings of WWW'2006, pp. 707–715 (2006)

[24] Rosso, M.A.: Using Genre To Improve Web Search. PhD thesis, University of North Carolina, Chapel Hill (2005)

[25] Russian Information Retrieval Evaluation Seminar (ROMIP), http://romip.ru

[26] Santini M.: Automatic Identification of Genre in Web Pages. PhD thesis, University of Brighton, UK (2007)

[27] Santini, M.: State-of-the-Art on Automatic Genre Identification. Technical Report ITRI-04-03, Information Technology Research Institute, Univ. of Brighton, UK (2004) Available online: ftp://ftp.itri.bton.ac.uk/reports/ITRI-04-03.pdf

[28] Si, L., Callan, J.: A Statistical Model for Scientific Readability. In: Proceedings of CIKM'2001, pp. 574–576 (2001)

[29] Strzalkowski, T., Guthrie, L., Karlgren, J., Leistensnider, J., Lin, F. Perez-Carballo, J., Straszheim, T., Wang, J., Wilding, J.: Natural Language Information Retrieval: TREC-5 Report. In: Proceedings of TREC'1995 (1996)

[30] Stubbe, A., Ringlstetter, Ch., Goebel, R.: Elements of a Learning Interface for Genre Qualified Search. In: Proceedings of the International Workshop "Towards Genre-Enabled Search Engines: The Impact of NLP", Borovets, Bulgaria, pp. 21–28 (2007)

[31] WEGA: Web Genre Analysis project, http://www.uni-weimar.de/cms/medien/webis/research/projects/wega.html